Heavy-Tailed Linear Bandits: Huber Regression with One-Pass Update

Jing Wang
LAMDA Group
School of Artificial Intelligence
Nanjing University

Outline

- Stochastic Linear Bandits
- Heavy-tailed Linear Bandits
- Our Results
- Conclusion

Outline

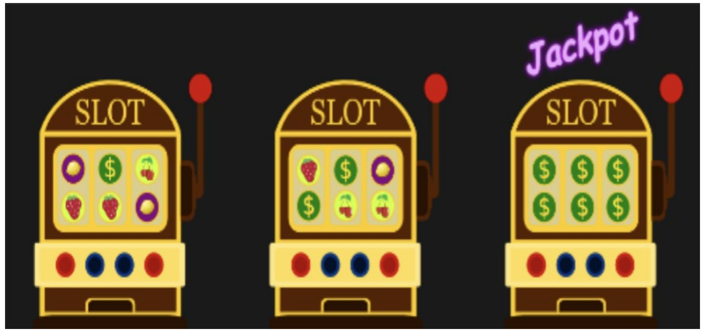
- Stochastic Linear Bandits
- Heavy-tailed Linear Bandits
- Our Results
- Conclusion

Stochastic Bandits

- Multi-Armed Bandits (MAB)

A player is facing K arms, and each time he pulls one arm and then receives a reward:

Arm 1	$X_{1,1}$	$X_{1,2}$	6	$X_{1,4}$	$X_{1,5}$
Arm 2	10	$X_{1,2}$	$X_{1,3}$	2	$X_{1,5}$
Arm 3	$X_{1,1}$	7	$X_{1,3}$	$X_{1,4}$	3



- Stochastic: rewards of the i-th arm are i.i.d. with unknown mean μ_i

Exploration vs **Exploitation**

- *Exploitation*: pull the best arm so far, for high reward
- *Exploration*: should try some other arms, they may be better

Stochastic Linear Bandits

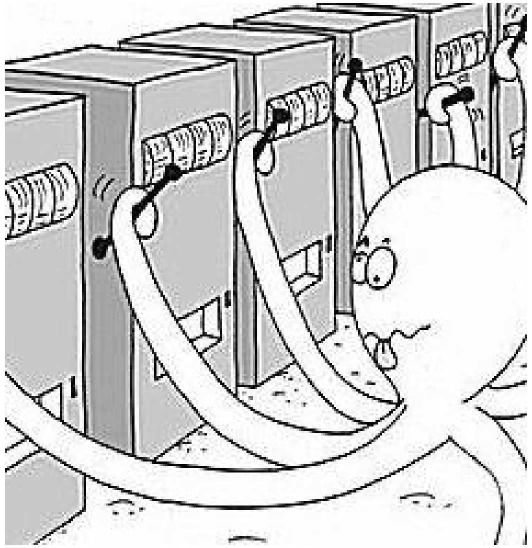
- Stochastic contextual bandit with a parametric model

Stochastic Linear Bandits

At each round $t = 1, 2, \cdots, T$

(1) the learner first chooses an arm $X_t \in \mathcal{X} \subseteq \mathbb{R}^d$;

(2) and then environment reveals a reward $r_t \in \mathbb{R}$.



➢ Linear reward model: $r_t = X_t^\top \theta_* + \eta_t$ **stochastic noise**

➢ Goal: minimize the regret $\text{REG}_T = \max_{\mathbf{x}\in\mathcal{X}} \sum_{t=1}^{T} \mathbf{x}^\top \theta_* - \sum_{t=1}^{T} X_t^\top \theta_*$

cumulative reward of the best offline model

Stochastic Linear Bandits (SLB)

LinUCB

for $t = 1$ to T **do**

Play X_t and observe reward r_t	**Step1. parameter estimation**
Parameter estimation $\widehat{\theta}_{t+1}$ of θ_* by **Least Squares**	$\widehat{\theta}_{t+1} = \arg\min_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{s=1}^{t} \left(X_s^\top \theta - r_s\right)^2$
Construct **Upper Confidence Bound** β_t	**Step2. arm selection**
Select $X_{t+1} = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \mathbf{x}^\top \widehat{\theta}_{t+1} + \beta_t \|\mathbf{x}\|_{V_t^{-1}} \right\}$	$\|\widehat{\theta}_{t+1} - \theta_*\|_{V_t} \leq \beta_t$ $\|\mathbf{x}\|_{V_{t-1}^{-1}}$: the degree of exploration of arm $\mathbf{x}$

end for

Theorem 1. *The regret of* LinUCB *is bounded with probability at least* $1 - 1/T$, *by*

$$\textsc{Reg}_T \leq \widetilde{\mathcal{O}}\left(d\sqrt{T}\right).$$

Further Application of SLB

Linear MDPs

$$r_h(x,a) = \langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}_h^* \rangle$$

$$\mathbb{P}_h(\cdot \mid x,a) = \langle \boldsymbol{\phi}(x,a), \boldsymbol{\mu}_h^*(\cdot) \rangle$$

- $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is known feature map
- $\{\boldsymbol{\theta}_h^*\}_{h=1}^H$ is the unknown reward parameter
- $\{\boldsymbol{\mu}_h^*(\cdot)\}_{h=1}^H$ is the unknown transition parameter

Algorithm: LSVI-UCB $\forall (x,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we have $Q_h^\pi(x,a) = \langle \phi(x,a), \mathbf{w}_h^\pi \rangle$.

Step1. Parameter estimation with Least Squares

$$\widehat{\mathbf{w}}_h = \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{\tau=1}^{k-1} \left[r_h\left(x_h^\tau, a_h^\tau\right) + \max_{a \in \mathcal{A}} Q_{h+1}\left(x_{h+1}^\tau, a\right) - \mathbf{w}^\top \boldsymbol{\phi}\left(x_h^\tau, a_h^\tau\right) \right]^2 + \lambda \|\mathbf{w}\|^2$$

Step2. Action selection with UCB $\quad a_h^k = \operatorname{argmax}_{a \in \mathcal{A}} \widehat{\mathbf{w}}_h^\top \boldsymbol{\phi}(x_h^k, a) + \beta \|\boldsymbol{\phi}(x_h^k, a)\|_{\Lambda_h^{-1}}$

Outline

- Stochastic Linear Bandits
- Heavy-tailed Linear Bandits
- Our Results
- Conclusion

Heavy-tailed Linear Bandits

- Linear reward with sub-Gaussian noise $r_t = X_t^\top \theta_* + \eta_t$

Assumption 1 (sub-Gaussian noise). The noise η_t is conditionally R-sub-Gaussian for some $R \geq 0$ i.e.

$$\forall \lambda \in \mathbb{R}, \mathbb{E}\left[\exp\left(\lambda \eta_t\right) \mid X_{1:t}, \eta_{1:t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

In many scenarios, the noise can be heavy-tailed !

- Linear bandits with heavy-tailed noise

Assumption 2 (heavy-tailed noise). The noise $\{\eta_t, \mathcal{F}_t\}$ is is martingale difference ($\mathbb{E}\left[\eta_t \mid \mathcal{F}_{t-1}\right] = 0$), and satisfies that for some $\varepsilon \in (0, 1], \nu_t > 0$,

$$\mathbb{E}\left[|\eta_t|^{1+\varepsilon} \,\middle|\, \mathcal{F}_{t-1}\right] \leq \nu_t^{1+\varepsilon}.$$

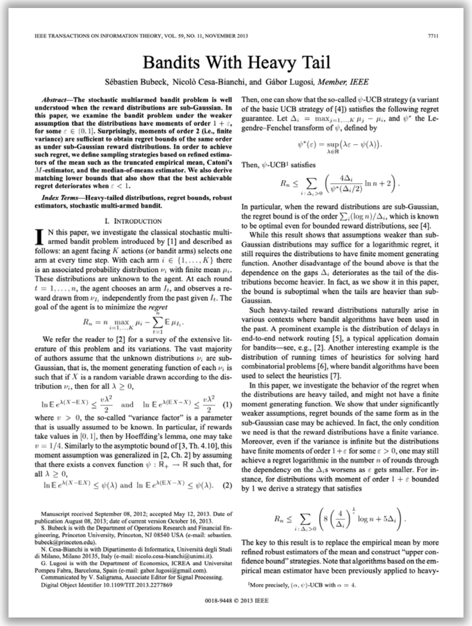
Challenge

Parameter estimation $\widehat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} \left(X_s^\top \theta - r_s\right)^2$

Key difficulty: the *large deviation* due to heavy-tailed noise.

Basic idea: reduce the impact of outliers

- ***Truncation***: directly removing data pair $\{X_s, r_s\}$ if r_s is extreme data;
- ***Median-of-Means***: repeat sampling same arm to reduce uncertainty;
- ***Robust loss function***: reduce penalty for large deviation $|X_s^\top \theta - r_s|$.


IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 59, NO. 11, NOVEMBER 2013 7711

Bandits With Heavy Tail

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi, *Member, IEEE*

Existing Methods

Limitation of truncation and Median-of-Means

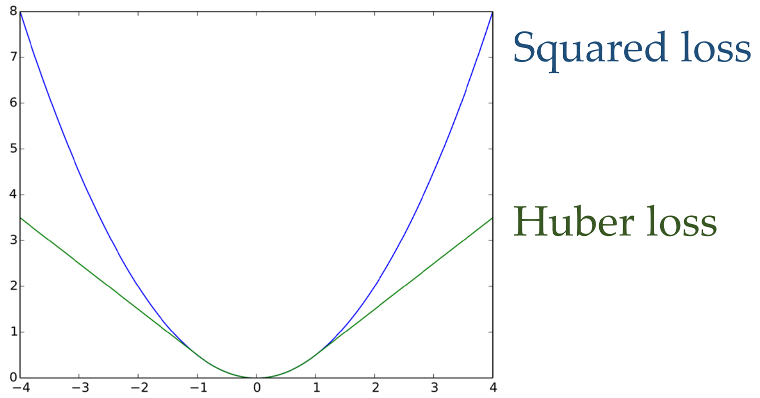
- Truncation relies on $\mathbb{E}\left[|r_t|^{1+\varepsilon} \mid \mathcal{F}_{t-1}\right] \leq u$, cannot recover noiseless case ☹
- Median-of-Means require repeated pulling and fixed arm set ☹

- Robust loss function

Definition 1 (Huber loss). Huber loss is defined as

$$f_\tau(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \leq \tau, \\ \tau|x| - \frac{\tau^2}{2} & \text{if } |x| > \tau, \end{cases}$$

where $\tau > 0$ is the robustification parameter.



- linear penalty for outliers
- squared penalty for inliers

Statistical Optimality

- HEAVY-OFUL Algorithm
  - ➢ ***Estimator***: adaptive Huber regression

$$\widehat{\theta}_t = \arg\min_{\theta \in \Theta} \frac{\lambda}{2}\|\theta\|_2^2 + \sum_{s=1}^{t-1} \ell_s(\theta)$$

  - ➢ ***Arm selection***: upper confidence bound

$$X_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \mathbf{x}^\top \widehat{\theta}_t + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}$$

With $z_s(\theta) = \frac{r_s - X_s^\top \theta}{\sigma_s}$, Huber loss is defined as

$$\ell_s(\theta) = \begin{cases} \frac{z_s(\theta)^2}{2} & \text{if } |z_s(\theta)| \leq \tau_s, \\ \tau_s |z_s(\theta)| - \frac{\tau_s^2}{2} & \text{if } |z_s(\theta)| > \tau_s. \end{cases}$$

Theorem 2. *The regret of* HEAVY-OFUL *is bounded with probability at least* $1 - 1/T$, *by*

$$\textsc{Reg}_T \leq \widetilde{\mathcal{O}}\left(dT^{\frac{1}{1+\varepsilon}}\right).$$

Efficiency Concern

- Adaptive Huber regression

$$\widehat{\theta}_t = \arg\min_{\theta\in\Theta} \frac{\lambda}{2}\|\theta\|_2^2 + \sum_{s=1}^{t} \ell_s(\theta)$$

The cost at round t

Computational cost: $\mathcal{O}(t \log T)$

Storage cost: $\mathcal{O}(t)$

- Least squares (closed-form solution)

$$\widehat{\theta}_t = V_{t-1}^{-1} \underbrace{\left(\sum_{s=1}^{t-1} r_s X_s\right)}_{Z_{t-1}}, V_{t-1} = \lambda I + \sum_{s=1}^{t-1} X_s X_s^\top$$

One-pass update

$$V_t = V_{t-1} + X_t X_t^\top$$

$$Z_t = Z_{t-1} + r_t X_t$$

Require one-pass algorithm for Heavy-tailed Linear Bandits !

Outline

- Stochastic Linear Bandits
- Heavy-tailed Linear Bandits
- Our Results
- Conclusion

Online Mirror Descent

- OMD is a powerful online learning framework to optimize regret.

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \eta_t \langle \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle + \mathcal{D}_\psi(\mathbf{x}, \mathbf{x}_t) \right\}$$

where $\mathcal{D}_\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$ is the Bregman divergence.

$$\widehat{\theta}_{t+1} = \arg\min_{\theta\in\Theta} \left\{ \left\langle \theta, \nabla \ell_t\left(\widehat{\theta}_t\right) \right\rangle + \mathcal{D}_{\psi_t}\left(\theta, \widehat{\theta}_t\right) \right\}$$

where $\psi_t(\theta) = \frac{1}{2}\|\theta\|_{V_t}^2$, with $V_t \triangleq \lambda I + \frac{1}{\alpha}\sum_{s=1}^{t} \frac{X_s X_s^\top}{\sigma_s^2}$

A Summary of OMD Deployment

- Our previous mentioned algorithms can all be covered by OMD.

Algo.	OMD/proximal form	$\psi(\cdot)$	η_t	Regret_T
OGD for convex	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2$	$\|\mathbf{x}\|_2^2$	$\frac{1}{\sqrt{t}}$	$\mathcal{O}(\sqrt{T})$
OGD for strongly c.	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_2^2$	$\|\mathbf{x}\|_2^2$	$\frac{1}{\sigma t}$	$\mathcal{O}(\frac{1}{\sigma}\log T)$
ONS for exp-concave	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \frac{1}{2}\|\mathbf{x}-\mathbf{x}_t\|_{A_t}^2$	$\|\mathbf{x}\|_{A_t}^2$	$\frac{1}{\gamma}$	$\mathcal{O}(\frac{d}{\gamma}\log T)$
Hedge for PEA	$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\Delta_N} \eta_t\langle \mathbf{x}, \nabla f_t(\mathbf{x}_t)\rangle + \text{KL}(\mathbf{x}\|\mathbf{x}_t)$	$\sum_{i=1}^{N} x_i \log x_i$	$\sqrt{\frac{\ln N}{T}}$	$\mathcal{O}(\sqrt{T\log N})$

Advanced Optimization (Fall 2024) Lecture 6. Online Mirror Descent 64

We here use OMD framework as a statistical estimation tool!

More details of OMD can be found in Lecture 6 of Advanced Optimization Course 2024 Fall
https://www.pengzhao-ml.com/course/AOpt2024fall/

Hvt-UCB

- OMD-based one-pass estimator

$$\widehat{\theta}_{t+1} = \arg\min_{\theta\in\Theta} \left\{ \left\langle \theta, \nabla \ell_t\left(\widehat{\theta}_t\right) \right\rangle + \mathcal{D}_{\psi_t}\left(\theta, \widehat{\theta}_t\right) \right\}$$

$$\psi_t(\theta) = \frac{1}{2}\|\theta\|_{V_t}^2 \text{ with } V_t \triangleq \lambda I + \frac{1}{\alpha}\sum_{s=1}^{t} \frac{X_s X_s^\top}{\sigma_s^2}$$

Computational Efficiency

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t - V_t^{-1}\nabla \ell_t\left(\widehat{\theta}_t\right)$$

$$\widehat{\theta}_{t+1} = \arg\min_{\theta\in\Theta} \left\| \theta - \widehat{\theta}_{t+1} \right\|_{V_t}$$

- Upper confidence bound

Lemma 1. *(Estimation error). If σ_t, τ_t, τ_0 are set as where $w_t \triangleq \frac{1}{\sqrt{\alpha}}\left\|\frac{X_t}{\sigma_t}\right\|_{V_{t-1}^{-1}}$ and let the step size $\alpha = 4$, then with probability at least $1 - 4\delta, \forall t \geq 1$, we have $\left\|\hat{\theta}_{t+1} - \theta_*\right\|_{V_t} \leq \beta_t$ with*

$$\beta_t \triangleq 107 \log \frac{2T^2}{\delta} \tau_0 t^{\frac{1-\varepsilon}{2(1+\varepsilon)}} + \sqrt{\lambda\left(2 + 4S^2\right)}, \quad \textit{where } \kappa \triangleq d \log\left(1 + \frac{L^2 T}{4\sigma_{\min}^2 \lambda d}\right).$$

Estimation Error

HEAVY-OFUL

MLE $\arg\min_{\theta\in\Theta} \frac{\lambda}{2}\|\theta\|_2^2 + \sum_{s=1}^{t} \ell_s(\theta)$

Estimation error $\widetilde{\mathcal{O}}\left(t^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)$

Hvt-UCB

OMD $\arg\min_{\theta\in\Theta} \left\{\left\langle\theta, \nabla\ell_t\left(\widehat{\theta}_t\right)\right\rangle + \mathcal{D}_{\psi_t}\left(\theta, \widehat{\theta}_t\right)\right\}$

Comp. cost per round $\mathcal{O}(1)$

Estimation error $\widetilde{\mathcal{O}}\left(t^{\frac{1-\epsilon}{2(1+\epsilon)}}\right)$

Theorem 3. *The regret of* Hvt-UCB *is bounded with probability at least* $1 - 1/T$, *by*

$$\text{REG}_T \leq \widetilde{\mathcal{O}}\left(dT^{\frac{1}{1+\varepsilon}}\right).$$

Instant-dependent Guarantee

- When ν_t is time-varying and known, Hvt-UCB can further achieve

Main Result

- Our work improves upon previous works **without additional assumptions**
  - *Statistical efficiency*: maintain the optimal and instant-dependent regret bound
  - *Computational efficiency*: reduce the per round time and storage cost

Method	Algorithm	Regret	Comp. cost	Remark
MOM	MENU [Shao et al., 2018] CRMM [Xue et al., 2023]	$\tilde{\mathcal{O}}\left(dT^{\frac{1}{1+\varepsilon}}\right)$	$\mathcal{O}(\log T)$ $\mathcal{O}(1)$	fixed arm set and repeated pulling
Truncation	TOFU [Shao et al., 2018] CRTM [Xue et al., 2023]	$\tilde{\mathcal{O}}\left(dT^{\frac{1}{1+\varepsilon}}\right)$	$\mathcal{O}(t)$ $\mathcal{O}(1)$	absolute moment $\mathbb{E}[\lvert r_t\rvert^{1+\varepsilon} \mid \mathcal{F}_{t-1}] \leq u$
Huber	HEAVY-OFUL [Huang et al., 2023]	$\tilde{\mathcal{O}}\left(dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}\sqrt{\sum_{t=1}^{T}\nu_t^2} + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}\right)$	$\mathcal{O}(t\log T)$	instance-dependent bound
Huber	Hvt-UCB (Corollary 1)	$\tilde{\mathcal{O}}\left(dT^{\frac{1}{1+\varepsilon}}\right)$	$\mathcal{O}(1)$	$\mathbb{E}[\lvert \eta_t\rvert^{1+\varepsilon} \mid \mathcal{F}_{t-1}] \leq \nu^{1+\varepsilon}$
Huber	Hvt-UCB (Theorem 1)	$\tilde{\mathcal{O}}\left(dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}\sqrt{\sum_{t=1}^{T}\nu_t^2} + dT^{\frac{1-\varepsilon}{2(1+\varepsilon)}}\right)$	$\mathcal{O}(1)$	instance-dependent bound

Estimation Error Analysis

- **Estimation error decomposition**

$$\left\|\widehat{\theta}_{t+1}-\theta_*\right\|_{V_t}^2 \leq \underbrace{2\sum_{s=1}^{t}\left\langle \nabla\tilde{\ell}_s\left(\widehat{\theta}_s\right)-\nabla\ell_s\left(\widehat{\theta}_s\right),\widehat{\theta}_s-\theta_*\right\rangle}_{\text{generalization gap term}}+\underbrace{\sum_{s=1}^{t}\left\|\nabla\ell_s\left(\widehat{\theta}_s\right)\right\|_{V_s^{-1}}^2}_{\text{stability term}}$$

Denoised loss: $\tilde{\ell}_t(\theta)=\frac{1}{2}\left(X_t^\top\left(\theta_*-\theta\right)/\sigma_t\right)^2$

$$+\underbrace{\left(\frac{1}{\alpha}-1\right)\sum_{s=1}^{t}\left\|\widehat{\theta}_s-\theta_*\right\|_{\frac{X_sX_s^\top}{\sigma_s^2}}^2}_{\text{negative term}}+4\lambda S$$

Ensure quadratic penalty for denoised data

$$\left|\left(X_t^\top\widehat{\theta}_t-X_t^\top\theta_*\right)/\sigma_t\right|\leq\frac{\tau_t}{2}$$

Recursive normalization factor tunning

$$\sigma_t=\max\left\{\nu_t,\sigma_{\min},\sqrt{\frac{2\beta_{t-1}}{\tau_0\sqrt{\alpha}t^{\frac{1-\varepsilon}{2(1+\varepsilon)}}}}\left\|X_t\right\|_{V_{t-1}^{-1}}\right\}$$

Estimation Error Analysis

- **Stability term** *Challenge of using one-pass OMD to approximate full-batch MLE*

$$\underbrace{2\sum_{s=1}^{t}\left(\min\left\{\left|\frac{\eta_s}{\sigma_s}\right|,\tau_s\right\}\right)^2\left\|\frac{X_s}{\sigma_s}\right\|_{V_s^{-1}}^2}_{\text{stochastic term}}+\underbrace{2\sum_{s=1}^{t}\left(\frac{X_s^\top\theta_*-X_s^\top\widehat{\theta}_s}{\sigma_s}\right)^2\left\|\frac{X_s}{\sigma_s}\right\|_{V_s^{-1}}^2}_{\text{deterministic term}}$$

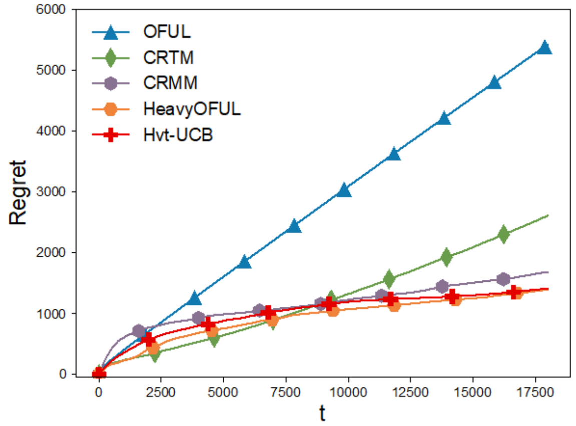
Concentration technique — *Canceled with negative term*

- **Generalization gap** *Challenge of handling Huber loss and heavy-tailed noise*

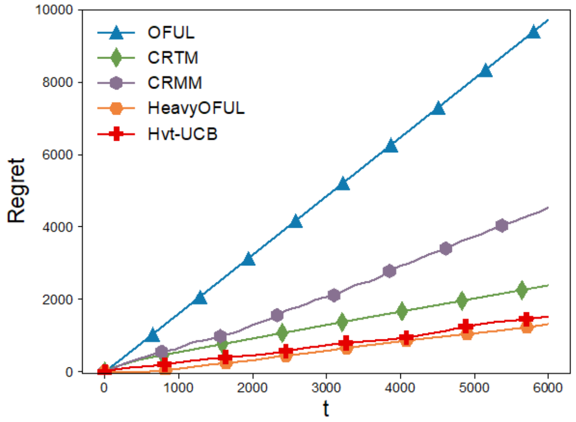
$$\underbrace{2\sum_{s=1}^{t}\left\langle\nabla\tilde{\ell}_s\left(\widehat{\theta}_s\right)+\nabla\ell_s\left(\theta_*\right)-\nabla\ell_s\left(\widehat{\theta}_s\right),\widehat{\theta}_s-\theta_*\right\rangle}_{\text{Huber-loss term}}+\underbrace{2\sum_{s=1}^{t}\left\langle-\nabla\ell_s\left(\theta_*\right),\widehat{\theta}_s-\theta_*\right\rangle}_{\text{self-normalized term}}$$

Concentration technique — *1-dimension self-normalized concentration*

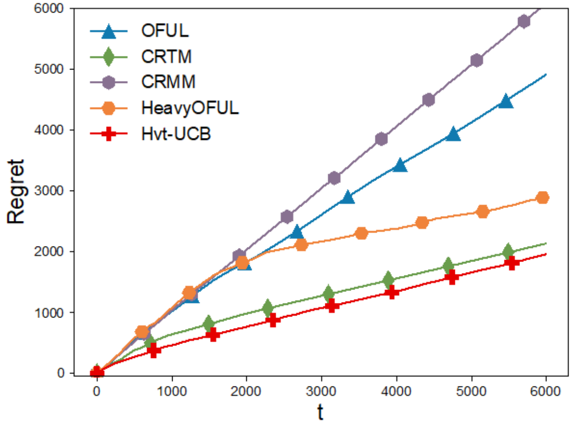
Experimental Results



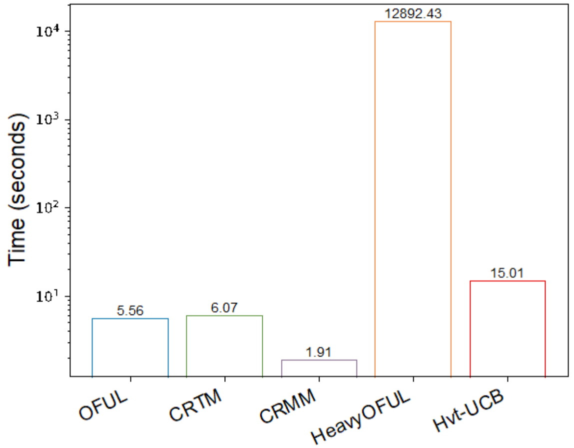
(a) Student t noise : regret



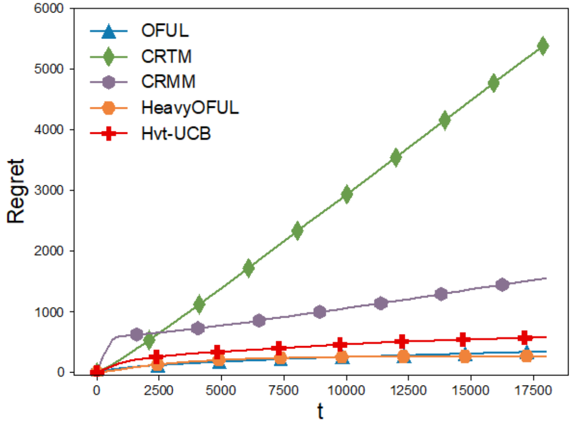
(c) Pareto noise: regret



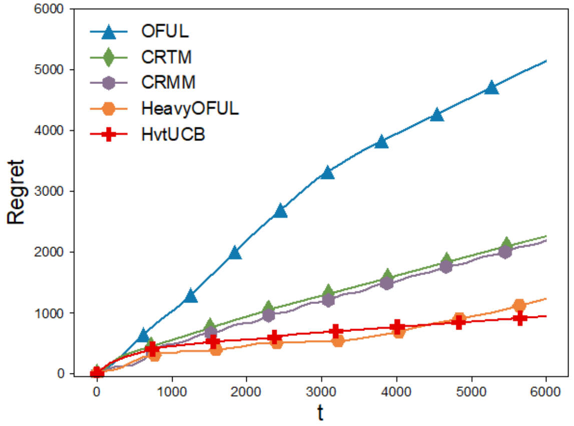
(e) Varying arm set



(b) Student t: running time



(d) Gaussian noise: regret



(f) Varying ν_t

Potential Extension

- Online Linear MDP

Realizable reward $R_h(s,a) = \langle \phi(s,a), \boldsymbol{\theta}_h^* \rangle + \epsilon_h(s,a)$

Known feature map

Heavy-tailed noise

Require: reward estimation under ***time-varying*** feature map

- Online Adaptive Control

State transition system $x_{t+1} = Ax_t + Bu_t + w_{t+1}$

Known state and action

Heavy-tailed noise

Require: system identification with ***finite-sample*** guarantee

Outline

- Stochastic Linear Bandits
- Heavy-tailed Linear Bandits
- Our Results
- Conclusion

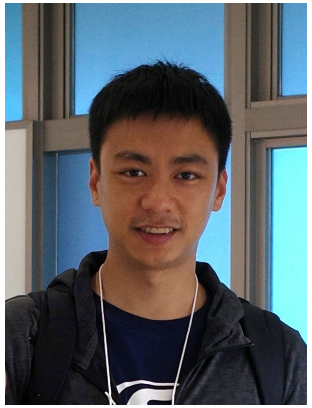
Conclusion

- **Problem**: heavy-tailed linear bandits
  - Only $(1+\epsilon)-$moment of noise is finite with $\epsilon \in (0,1)$
- **Approach**: Huber loss-based one-pass algorithm
  - Employing OMD with tailored local norm to replace the MLE in SLB
  - Achieve the optimal and variance-aware regret bound with O(1) cost

Open Questions

- How to handle unknown variance ν_t while maintaining current guarantee

Joint work with



Yu-Jie Zhang (RIKEN AIP)

Peng Zhao (NJU)

Zhi-Hua Zhou (NJU)

Jing Wang, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Heavy-Tailed Linear Bandits: Huber Regression with One-Pass Update, ICML2025.

Thanks!

Other References

- Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. COLT 2011.
- Yasin Abbasi-yadkori, Dávid Pál, Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. NIPS 2011.
- Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. TIT 2013.
- Han Shao, Xiaotian Yu, Irwin King, and Michael R. Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. NeurIPS 2018
- Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. JASA 2020
- Y.-J. Zhang and M. Sugiyama. Online (Multinomial) Logistic Bandit: Improved Regret and Constant Computation Cost. NeurIPS 2023.
- Jiayi Huang, Han Zhong, Liwei Wang, and Lin Yang. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. NeurIPS 2023
- Bo Xue, Yimu Wang, Yuanyu Wan, Jinfeng Yi, and Lijun Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. NeurIPS 2023.